Student's Semester Examination Knowledge Gap Categorization Using Classification Techniques

Kaustuv Deb*
Department of Computer Science & Engineering, Supreme Knowledge Foundation Group of Institutions, Chandannagar, India
kaustuvdeb24@gmail.com

Debabrata Jana*
Department of Master of Computer Applications, Calcutta Institute of Technology, Howrah, India
debabratajana03@gmail.com

Atanu Das*
Department of Master of Computer Applications, Netaji Subhash Engineering College, Kolkata, India
atanudas75@gmail.com

Rajib Bag*
Principal, Indas Mahavidyalaya, Bankura, India
rajib.bag@gmail.com

* *Corresponding author*

Abstract:

Categorization of the knowledge gap present in each student and accordingly applying the appropriate remedy to remove the knowledge gap is essential in education system to ensure that the student gains proper knowledge. Performance of a student is evaluated through a number of internal examinations and an end semester examination. Categorization of the semester examination knowledge gap of a student is possible through a proper and tedious analysis of the internal examination knowledge deficiencies of the student. This paper proposes the development of an automatic machine learning based system for accurately categorizing the semester examination knowledge gap of a student by analyzing the internal examination knowledge deficiencies of the student thoroughly using J48, Random Forest and Naïve Bayes classification techniques. The categorization of each student's future end semester examination knowledge gap in advance will be very much useful for providing proper advice to each student to fill up the student's knowledge gap and upgrade student's performance. Results produced by the proposed system advocate that the proposed system has performed the required categorization task perfectly. Classifiers are applied using 10-fold cross-validation technique. Random Forest classifier has produced 72.02% accuracy. J48 and Naïve Bayes classifier has produced 71.63% and 66.47% accuracy respectively.

Keywords: Knowledge gap categorization, classification techniques, J48 classifier, Random Forest classifier, Naïve Bayes classifier, performance evaluation, knowledge deficiency.

Introduction

Performance evaluation of students is a very important part of any education system. The evaluation process should be perfect enough for finding gaps of knowledge existing in students. Categorizing knowledge gaps and reducing it as much as possible with proper remedies is highly crucial in any education system to make the students able to get the proper benefits of learning. In general, performance of each student, in a particular semester, is monitored by the human tutors through the results of different internal examinations and an end semester examination. Tutorial classes are provided to take care of the gap of knowledge existing in each student. But proper categorization of the semester examination knowledge gap existing in each student through a thorough and perfect analysis of each student's performances in the internal examinations is essential for identifying the required tutorial classes. Knowledge gap categorization is a very tedious, complicated and error prone task for the human tutors to perform and requires great deal of experience to accomplish it perfectly. An automated system, capable of carrying out this knowledge gap categorization task perfectly, can be very much beneficial for the education systems. Machine learning based classification techniques like: J48, Random Forest and Naive Bayes are widely used for accomplishing various classification activities in educational fields. These simple and efficient classifiers can analyze educational data vividly through complicated calculations without any error and can make classification decisions with great deal of accuracy in quick time. In this paper, development of an automated machine learning based system to categorize the semester examination knowledge gap existing in each student through a vivid analysis of each student's internal examination knowledge deficiencies applying J48, Random Forest and Naïve Bayes classifiers, is proposed. Each student's knowledge deficiencies in three internal examinations are considered and analyzed by the proposed system towards achieving the goal of end semester knowledge gap categorization of each student. A brief introduction of other sections of this paper is given below-

Section II presents literature survey, section III portrays the classification techniques used in this paper, section IV describes the proposed system, section V showcases the experimental results and finally the conclusion of the proposed work is drawn in section VI.

Literature Review

Several research activities have been done for student's performance analysis and student's deficiency identification using various data mining and soft computing techniques. Authors in [1] carried their research work in predicting performance of students using decision tree classifier. This performance analysis was useful for providing guidance to students for improvement. In [2] prediction of student performance using classifiers such as J48 decision tree, Naïve Bayes, Bayesian Network, k-Nearest Neighbour, OneR, JRip was performed and the highest prediction accuracy was produced by J48 classifier. A system for analyzing result of university students and performing prediction using J48, REPTree, and Hoeffding Tree decision tree classifiers was developed in [3] and J48 classifier achieved the highest prediction accuracy among three applied classifiers. This research focused on giving suggestions to students to overcome their drawbacks to enhance their performance. Authors in [4] have done a vivid review on applications of data mining in education sector. This review portrayed that applications of data mining techniques in education sector provide enormous benefits to teachers, students and educational institutions such as properly predicting performance of students, providing detailed feedback for improving teaching-learning process and quality enhancement of educational institutions, identifying weak students for proving special care for their performance improvement etc. Thus, data mining plays a very important role in upgrading education system. Authors in [5] proposed a model for predicting final marks of students based on data in a forum using classification via clustering. This research found that participation of students in the forum of course can predict the final marks of students well. Authors in [6] proposed a method for identifying tutorial gap of student through vivid analysis of MCQ test responses of students. In this research MCQ questions related to subtopics taught to students were developed and the answers given by each student were analyzed properly for finding out the tutorial gap existing in the student at subtopic level. Authors in [7] developed a data mining based model for predicting academic performance of students using ensemble methods. Here behavioral attributes of students are used in prediction. This model was beneficial for education system since proper understanding of students, identification of weak students and up gradation of method of learning was possible using this model. Authors in [8] presented a method for student's tutorial gap identification applying fuzzy logic. Each student was first assessed through MCQ test and verbal explanation at subtopic level and then each student's responses were analyzed using a properly developed fuzzy inference system to identify tutorial gap present in each student efficiently. This research proposed a realistic fuzzy logic blended approach for the identification of actual lacking of each student in the depth of taught subtopic. Authors in [9] compared different supervised data mining methods in terms of prediction of exam performance of student and found artificial neural network achieved the highest precision.

Classification Technique Used

A. J48 Classifier

J48 is a decision tree classifier. This classifier is very simple and easy to use. This classifier can produce excellent classification results when applied on a data set consisting of instances where each instance is described by a set of features. This classifier has a tree structure where features of the data set instances are represented by non-leaf nodes and classes of the data set instances are represented by leaf nodes. Internal node gets split into internal nodes or leaf nodes and thereby the decision tree grows. When all branches of the decision tree produce leaf nodes, splitting stops. Classes of data set instances are found when leaf nodes are reached. If a single leaf node is produced from the feature values of more than one instances, then those all instances are assigned same class label.

B. Random Forest Classifier

Multiple decision trees are built in Random Forest classification technique. Thereafter, these trees are merged for generating prediction output with high accuracy. Here, combined prediction output is obtained which is better than the prediction output of a single decision tree. As this classifier produces combined prediction output, this classifier can be trusted highly.

C. Naïve Bayes Classifier

Naïve Bayes classifier is developed depending on Bayes theorem. Here, features of the data set are needed to be independent. The prediction output is generated by the contribution of each independent feature. This probability based classifier is a very strong classifier and produces accurate classification outputs.

[2] Proposed System

The proposed system categorizes the semester examination knowledge gap of each student in the following way-

At first, marks of three internal examinations of 550 students, studying in Computer Science Engineering, for the subject Database Management System, are collected and internal examination knowledge deficiency, of each student for each of the three internal examinations, is obtained from the collected marks. Internal examination knowledge deficiency can take any of the three values such as Minor (for internal examination marks 71% to 99%), Major (for internal examination marks 31% to 70 %) and Extreme (for internal examination marks 0% to 30%). Internal examination knowledge deficiencies of a student effectively lead to the categorization of the future end semester examination knowledge gap of the student. Thus, the proposed system uses knowledge deficiencies of three internal examinations of a student as predictors for predicting the appropriate category of the semester examination knowledge gap of the student. After data collection, the collected data set is pre-processed properly by eliminating the missing values and a pre-processed data set having a size of 504 is obtained. Preprocessing is carried out to ensure that the dataset becomes perfectly suited for the application of the classifiers. After preprocessing, classifiers such as J48, Random Forest and Naïve Bayes are applied one by one on the pre-processed data set using 10-fold cross-validation technique for categorizing each student's semester examination knowledge gap. Semester examination knowledge gap is categorized into three classes such as Small (for semester examination marks 71% to 99%), Avg (for semester examination marks 31% to 70 %) and Large (for semester examination marks 0% to 30 %). Weka, the most widely used software in the field of machine learning, is used to apply the classifiers easily without any hazard. Figure I shows the block diagram of the proposed system.

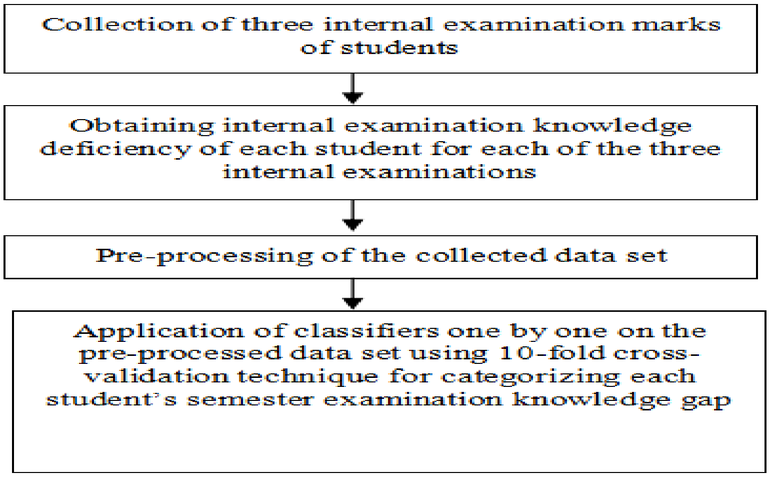
Collection of three internal examination marks of students
Obtaining internal examination knowledge deficiency of each student for each of the three internal examinations
Pre-processing of the collected data set
Application of classifiers one by one on the pre-processed data set using 10-fold cross-validation technique for categorizing each student's semester examination knowledge gap

Figure I. Block diagram of the proposed system

Experimental results and Discussion:

Performance of each classifier is evaluated using some very popular parameters such as Accuracy, True Positive Rate (TP Rate) or Recall, False Positive Rate (FP Rate), Precision and F-Measure. Said parameters are defined below-

Accuracy: It is defined as the percentage of data set instances classified correctly. Accuracy is calculated using equation (1).

Accuracy=((Number of data set instances classified correctly) / (Total number of data set instances)) X 100% (1)

TP Rate or Recall: True positive instances of a class are the number of instances those actually belong to the class and are also correctly classified by the classifier to belong to the class. False negative instances of a class are the number of instances those actually belong to the class and are wrongly classified by the classifier to belong to other class. Ratio of the true positive instances of the class and the total number of the true positive instances and false negative instances of the class defines the TP Rate or Recall of a class. TP Rate or Recall of a class is calculated using equation (2).

TP Rate or Recall=(True positive instances of the class) / (True positive instances of the class + False negative instances of the class) (2)

FP Rate: FP rate of a class is the ratio of the false positive instances of the class and the total number of the false positive instances and true negative instances of the class. False positive instances of a class are the number of instances those actually don’t belong to the class and are wrongly classified by the classifier to belong to the class. True negative instances of a class are the number of instances those actually don’t belong to the class and are correctly classified by the classifier as not belonging to the class. FP Rate of a class is calculated using equation (3).

FP Rate=(False positive instances of the class) / (False positive instances of the class + True negative instances of the class) (3)

Precision: Precision of a class is the ratio of the true positive instances of the class and the total number of the true positive instances and false positive instances of the class. Precision of a class is calculated using equation (4).

Precision=(True positive instances of the class) / (True positive instances of the class + False positive instances of the class) (4)

F-Measure: F-Measure of a class is twice of the ratio of the product of Precision of the class and Recall of the class and sum of Precision of the class and Recall of the class. F-Measure of a class is calculated using equation (5).

F-Measure=((Precision of the class X Recall of the class) / (Precision of the class + Recall of the class)) X 2 (5)

Confusion Matrix is an important and useful tabular form for depicting the performance of a classifier. Confusion matrix is defined below-

Confusion Matrix: Confusion Matrix is a matrix that shows the performance of a classifier. Each row of a Confusion Matrix depicts the instances actually belonging to a class and each column of a Confusion Matrix depicts the instances classified by the classifier to belong to a class.

Tables I, III and V show the classification performances of the applied classifiers for each class of semester examination knowledge gap in terms of Accuracy, TP Rate or Recall, FP Rate, Precision and F-Measure. Tables II, IV and VI show the Confusion Matrices of the applied classifiers.

Table I shows the performance of J48 classifier.

Class	Accuracy	TP Rate or Recall	FP Rate	Precision	F-Measure
Small	71.63 %	0.623	0.068	0.745	0.679
Avg		0.843	0.284	0.657	0.739
Large		0.641	0.094	0.797	0.711

Table I. Performance Of J48 Classifier

Table II shows the Confusion Matrix of J48 classifier.

Small	Avg	Large	Classified as

76	35	11	**Small**
12	167	19	**Avg**
14	52	118	**Large**

Table II. Confusion matrix of j48 classifier

Table III shows the performance of Random Forest classifier.

Class	Accuracy	TP Rate or Recall	FP Rate	Precision	F-Measure
Small	72.02%	0.623	0.065	0.752	0.682
Avg		0.854	0.297	0.650	0.738
Large		0.641	0.078	0.825	0.722

Table III. Performance of random forest classifier

Table IV shows the Confusion Matrix of Random Forest classifier.

Small	Avg	Large	Classified as
76	37	9	**Small**
13	169	16	**Avg**
12	54	118	**Large**

Table IV. Confusion Matrix Of Random Forest Classifier

Table V shows the performance of Naïve Bayes classifier.

Class	Accuracy	TP Rate or Recall	FP Rate	Precision	F-Measure
Small	66.47%	0.631	0.128	0.611	0.621
Avg		0.657	0.196	0.684	0.670
Large		0.696	0.188	0.681	0.688

Table V. Performance Of Naïve Bayes Classifier

Table VI shows the Confusion Matrix of Naïve Bayes classifier.

Small	Avg	Large	Classified as
77	21	24	**Small**

32	130	36	**Avg**
17	39	128	**Large**

Table VI. Confusion matrix of naïve bayes classifier

Table I, III and V portray that Random Forest classifier has produced the highest classification accuracy of 72.02% . J48 classifier and Naïve Bayes classifiers also have exhibited good deal of accuracies. Highest TP Rate or Recall 0.631 for class Small is produced by Naïve Bayes classifier. Highest TP Rate or Recall 0.854 for class Avg is produced by Random Forest classifier. Highest TP Rate or Recall 0.696 for class Large is produced by Naïve Bayes classifier. Lowest FP Rate 0.065 for class Small is produced by Random Forest classifier. Lowest FP Rate 0.196 for class Avg is produced by Naïve Bayes classifier. Lowest FP Rate 0.078 for class Large is produced by Random Forest classifier. Highest Precision of 0.752 for class Small is produced by Random Forest classifier. Highest Precision of 0.684 for class Avg is produced by Naïve Bayes classifier. Highest Precision of 0.825 for class Large is produced by Random Forest classifier. Highest F-Measure of 0.682 for class Small is produced by Random Forest classifier. Highest F-Measure of 0.739 for class Avg is produced by J48 classifier. Highest F-Measure of 0.722 for class Large is produced by Random Forest classifier. No classifier has produced adverse results like low TP Rates or Recall, high FP Rates, low Precisions or low F-Measures, rather all classifiers have produced high TP Rates or Recall, low FP Rates, high Precisions and high F-Measures.

Conclusion

Results produced by the proposed system establish the fact that the proposed system has properly accomplished the task of categorization of each student's semester examination knowledge gap. The proposed system has not produced any adverse result; hence the proposed system is trustworthy. The proposed system can be beneficial for the teachers and the students since the proposed system can predict the future end semester examination knowledge gap of each student correctly and thereby the proposed system can provide the scope of correction to each student for upgrading student's performance. The proposed system, being applied on the data set of engineering students, has exhibited notable performance. In the future the proposed system is to be tested on the data sets of the students belonging to other courses. Also other classifiers like: ANN and SVM may be incorporated in the proposed system in the future research.

References

[3] Raut, A. B., & Nichat, M. A. A. (2017). Students performance prediction using decision tree. *International Journal of Computational Intelligence Research, 13*(7), 1735-1741.

[4] Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies, 13*(1), 61-72.

[5] Hoque, M. I., Kalam Azad, A., Tuhin, M. A. H., & Salehin, Z. U. (2020). University students result analysis and prediction system by decision tree algorithm. *Adv Sci Technol Eng Syst J, 5*(3), 115-122.

[6] Guleria, P., & Sood, M. (2014). Data mining in education: a review on the knowledge discovery perspective. *International Journal of Data Mining & Knowledge Management Process, 4*(5), 47-60.

[7] Lopez, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. *International Educational Data Mining Society*, 148-151.

[8] Das, A., Deb, K., Bajerjee, S., & Bag, R. (2017). A new method for tutorial gap identification towards students modeling. *Mathematical Modelling of Engineering Problems, 4*(2), 80-83.

[9] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136.

[10] Deb, K., Banerjee, S., Das, A., & Bag, R. (2019). Tutorial gap identification towards student modeling using fuzzy logic. *International Journal of Information and Communication Technology Education (IJICTE), 15*(3), 30-41.

[11] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & education, 143,* 103676.